图书情报工作
第63卷 第10期 2019年5月

基于网络招聘文本挖掘的课程知识模型自动构建研究*

俞琰[1,2] 陈磊[1] 赵乃瑄[1]
[1] 南京工业大学信息服务部 南京 210009 [2] 东南大学成贤学院计算机工程系 南京 211816

摘要：［**目的/意义**］为帮助高校师生充分利用网络招聘信息，提出基于大数据量网络招聘文本挖掘的课程知识模型及其自动构建方法。［**方法/过程**］本文提出包含“岗位－课程－知识点”的三级课程知识模型，利用自然语言文本挖掘技术实现课程知识点模型的自动构建，并通过实验对其构建过程进行验证和分析。［**结果/结论**］实验结果表明本文提出的模型及方法具有高度的可行性与有效性，可为高校和学生提供教学和学习参考。

关键词：网络招聘文本 课程知识模型 文本挖掘
分类号：G202
DOI:10.13266/j.issn.0252－3116.2019.10.015

近几年来，随着我国高等教育的迅猛发展和招生规模的日益扩大，大学生找工作难、企业招人难已经成为社会关注的热点。在某种程度上，我国高校人才培养与社会需求间的不匹配，造成了这种双重困境。特别地，在信息时代中，企业对人才的需求变化迅速，与之相矛盾的是高校人才培养周期长，专业课程设置滞后，导致学生的培养脱离实际需要。因此，在高速发展的信息时代中，快速、准确地洞察企业对所招岗位的知识需求显得格外重要。随着互联网的普及，网络招聘成为企业招聘的主流方式。网络招聘文本中常含有企业对所招岗位专业知识需求的具体描述，反映了当前就业市场对人才的专业知识需求。因此，网络招聘文本分析是了解整个社会对某领域人才知识需求的一种有效途径。

虽然一些学者已经意识到其重要性并开展网络招聘文本分析的研究，但是目前研究还存在如下两个主要问题：①研究主要对岗位所需技能知识进行统计，没有进一步利用网络招聘文本信息；②分析主要以手工方法为主，不能满足大数据时代招聘网络数据量大、变化快速的要求。

针对目前研究存在的问题，本文提出一个包含“岗位－课程－知识点”的课程知识模型，并利用文本挖掘技术，自动构建课程知识模型，以适应大数据时代数据量大、数据变化快速的特点。最后，对计算机相关专业的网络招聘文本进行实证分析。实证结果表明本模型以及构建过程的可行性与有效性。课程知识模型可以帮助高校根据社会对特定领域人才技能的需求，不断优化专业课程体系与教学大纲，为其制定符合企业需求的专业人才培养方案提供情报决策支持。课程知识模型还可以帮助学生根据自身兴趣与欲从事的岗位，有重点地加强某些专业课程及其知识点的学习。

1 相关研究

网络招聘文本分析通常包括招聘实体信息抽取与招聘实体分析两个步骤。

招聘实体信息抽取是指从半结构化的网络招聘文本中抽取结构化的招聘实体信息，如岗位、技能、专业等信息。根据抽取方法的不同，可分为手工方法和自动方法两大类。手工方法直接人工抽取网络招聘文本中岗位、所需技能等信息。如：C. Chao 和 S. Shih[1]采集 Monster 招聘网站信息，手工抽取招聘岗位、技能等信息；I. Wowczko[2]手工抽取和映射招聘中的技能；J. Y. Kim 和 C. K. Lee[3]手工分析数据科学家招聘信息；D. A. Mauro 等[4]手工抽取工作类型所需的技能；

* 本文系教育部人文社会科学规划项目“大数据时代技能知识图谱构建研究”（项目编号：16YJAZH073）和国家社会科学基金一般规划项目“大数据时代支持创新设计的多维度多层次专利文本挖掘研究”（项目编号：17BTQ059）研究成果之一。

作者简介：俞琰（ORCID：0000－0002－9654－8614），副教授，博士，E-mail：yuyanyuyan2004@126.com；陈磊，硕士研究生；赵乃瑄（ORCID：0000－0001－9072－7315），教授。

收稿日期：2018－08－30 修回日期：2018－12－03 本文起止页码：134－142 本文责任编辑：王传清

吕斌等[5]、李国秋[6]调研300个情报职业招聘网页，手工抽取情报职业要求、职业类型、职责和作用等信息；夏火松和潘筱听[7]手工抽取我国大数据企业人才需求信息；黄崑等[8]从网络招聘文本手工抽取图情岗位对人才岗位、知识和能力的要求信息。贾东琴和檀博[9]手工抽取 ALA Joblist、IFLA 的 LIBJOBS Mailing List 以及 ACRL3 个招聘网站文本中的招聘实体。

显然，手工方法很难满足大数据量、非结构化环境下的网络招聘信息分析要求。一些研究尝试使用基于外部资源、基于规则、基于统计、基于深度学习的方法自动抽取网络招聘文本中的信息。基于外部资源的方法利用技能词典、维基百科等资源，构建专业知识词典以抽取信息。如：M. Sodhi 和 B. Son[10]构建运筹学专业核心词典。M. Zhao 等[11]使用常规短语、领域专家预定义的各种术语分析招聘网页。T. Xu[12]等从 CSDN 网站下载技能种类和具体技能，共 54 个技能种类和 1 729 个具体技能，构建了专业知识字典。詹川[13]参考已有的电子商务专业知识，构建该专业的术语词典，从招聘文本中抽取高于一定频数的技能。夏立新等[14]利用中华教育在线职业大全、招聘网岗位分类、论文关键词构建专业、岗位和知识点词典，抽取专业、岗位、技能等信息。然而，基于外部资源的方法存在外部资源更新较慢、覆盖面较窄的问题。基于规则的方法人工构造规则模板，以实现信息抽取。如：M. Bastian 等[15]利用逗号进行匹配，抽取 LinkedIn 网络招聘文本中的技能信息。王召义等[16]使用具备、熟悉、精通、能力这 4 个词作为邻近词，构建抽取规则，以抽取岗位所需的技能。基于规则的方法存在方法过于简单、结果不尽理想等问题。基于统计的方法主要利用语料库统计某个词的概率信息，以抽取招聘实体。如：刘睿伦等[17]采用词频统计信息抽取招聘实体。张俊峰和魏瑞斌[18]抓取前程无忧、智联招聘等专业招聘网站数据，使用词频等方法抽取招聘实体，以构建招聘词典。基于统计的方法也存在方法过于简单，结果不尽理想等问题。随着深度学习的迅速发展，王东波等[19]利用深度学习模型，设计数据科学招聘实体自动抽取平台。然而，深度学习方法需要大规模人工标注语料作为训练数据，目前网络招聘技能信息抽取任务没有大规模标注语料库。

网络实体信息分析是指对抽取的结构化招聘实体信息进行分析的过程。目前的分析主要是对抽取的岗位、技能、专业等信息进行统计，没有充分地利用网络招聘文本信息。如，J. Y. Kim 等[3]分析数据科学家招聘信息，总结企业对数学专业及学历要求。D. A. Mauro 等[4]结合专家判断，分析 2 700 条大数据相关岗位信息，对每一个工作类型所需的技能和熟练程度要求进行评估。吕斌等[5]、李国秋[6]调研 300 个情报职业招聘网页，分析社会组织的情报职业需求，以及社会组织中情报职业类型、职责和作用等。黄崑等[8]从职位基本信息、岗位职责、任职要求 3 个角度分析大数据岗位对人才知识和能力的要求。M. Sodhi 和 B. Son[10]以研究不同行业对运筹专业技能需求的差异。詹川[13]分析电商各岗位的需求、技能整体需求和各个岗位特别需求的技能。魏来和郑华敏[20]对国内外高校图书馆招聘信息进行调研，从统计知识背景、综合素质、岗位职责、职业技能和特殊技能 5 个方面剖析数据馆员需要具备的职业能力。贾东琴和檀博[9]分析了国外高校图书馆岗位需求中岗位数量、岗位职责要求、入门资质、加分资质等要求。田野[21]针对 2016 年度 1 359家机构的图情专业招聘需求数据，从招聘机构类型、人数、地域、招聘对象学历要求、岗位偏好等方面进行了实证分析。陈媛媛和董伟[22]借助社会网络分析工具对招聘广告中的就业技能及其关系进行研究。

2 课程知识模型逻辑结构

目前研究主要对岗位所需技能知识点进行统计，如图 1(a)所示，没有进一步利用网络招聘文本信息。针对这个问题，本文提出包含“岗位 - 课程 - 知识点”的三级课程知识模型，如图 1(b)所示：

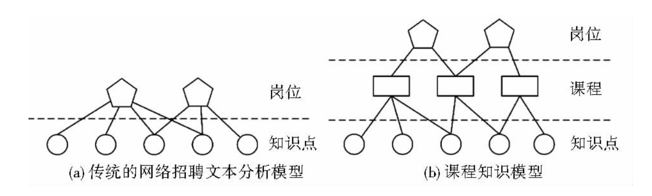


图 1 网络招聘文本挖掘

课程知识模型包含岗位、课程、知识 3 个对象。其中，岗位为企业要求员工完成的一项或多项责任以及为此赋予员工的权力的总和；课程指高校根据培养目标所开设的专业知识和专门技能的课程；知识点为岗位所需要的知识以及专业技能，也是课程包含知识的基本单元。课程知识模型还包括岗位 - 课程、课程 - 知识两种关系。其中，岗位和课程之间存在多对多关系，即一个岗位需要学习若干门课程，一门课程可应用于若干个相关岗位；课程与知识点之间也存在多对多关系，即一门课程包含若干知识点，一个知识点也可归

属于若干门课程。图 2 以“大数据工程师”岗位为例，表明针对该岗位应学习的主要课程，以及课程所包括的主要知识点，其中对象间连线的粗细表示对象关系的强弱。

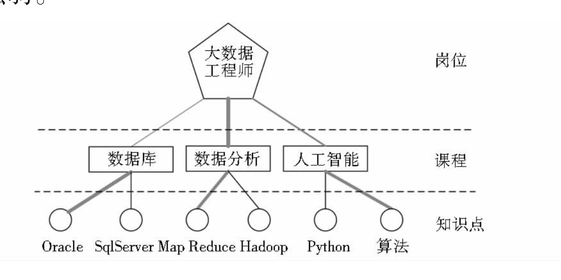


图 2　课程知识模型具体示例

3　课程知识模型自动构建

网络招聘文本通常包含岗位、岗位职责、任职要求等信息，图 3 为一个网络招聘文本示例。图 3(a)为网络招聘文本在浏览器中的显示页面，图 3(b)为对应的 HTML 文本。岗位描述了企业招人所从事的岗位名称；岗位职责描述了该岗位需要承担的责任；任职要求描述了该岗位人员应该具备的专业知识技能以及其他基本能力。

课程知识模型中的岗位信息可以从招聘网络文本中的岗位部分直接抽取，知识点可以从任职要求对应的文本中抽取。课程信息使用主题模型生成，并根据

Java 开发工程师

岗位职责：

1.参与代码设计、审核、检查；

2.能独立解决开发中遇到的疑难问题；

3.完成核心、重要模块的设计、开发、测试；

4.参与系统稳定性、扩展性、性能调试。

任职要求：

1.1 年以上 Java 开发经验，对软件工程和相关标准有良好的认识，具有较强的面向对象思维；精通设计模式；

2.熟悉 Spring、MyBatis　等主流 J2EE　技术；熟练使用 Oracle 数据库，并有一定的 SQL 优化经验；

3.熟悉 Javascript、JQuery、BootStrap、CSS 等技术；熟悉 Linux　操作系统；熟悉 Tomcat　应用服务器；

4.有 Spark、Hadoop　开发经验者优先；

5.能够承受压力、基础扎实、思路清晰，有独立解决问题的能力、良好的沟通表达能力，有责任心，具有良好的团队合作意识。

（a）招聘网络文本显示页面

<div　class="tHeader　tHjob"><div　class="in"><div　class="cn"><h1 title="Java 开发工程师">Java 开发工程师</h1><div　class="bmsg job_msg inbox"><p>岗位职责：</p><p>1.参与代码设计、审核、检查；</p><p>2.能独立解决开发中遇到的疑难问题；</p><p>3.完成核心、重要模块的设计、开发、测试；</p><p>4.参与系统稳定性、扩展性、性能调试。</p><p>任职要求：</p><p>1.1 年以上 Java 开发经验，对软件工程和相关标准有良好的认识，具有较强的面向对象思维；精通设计模式；　</p><p>2.熟悉 Spring、MyBatis 等主流 J2EE 技术；熟练使用 Oracle　数据库，并有一定的　SQL　优化经验；</p><p>3.熟悉 Javascript、JQuery、BootStrap、CSS 等技术；熟悉 Linux 操作系统；熟悉 Tomcat 应用服务器；</p><p>4.有 Spark、Hadoop 开发经验者优先；</p><p>5.能够承受压力、基础扎实、思路清晰，有独立解决问题的能力、良好的沟通表达能力，有责任心，具有良好的团队合作意识。</p>

（b）招聘网络文本 HTML 页面

图 3　网络招聘文本示例

主题模型和统计信息生成岗位－课程关系和课程－知识点关系。因此，本文提出的课程知识模型构建流程见图 4，主要包括数据抓取、岗位抽取、知识点抽取、课程生成、岗位－课程关系生成、课程－知识点关系生成等 6 个步骤。

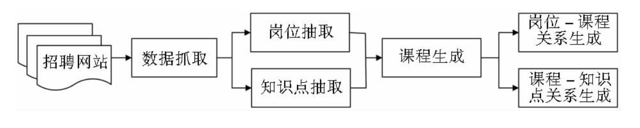


图 4　课程知识模型构建流程

3.1　数据抓取

选择合适的招聘网站，选取相关专业，使用 Python 脚本，首先获取网络招聘文本 URL，并将其推送到 Django Rest 接口收集端，然后根据收集端职位 URL，逐一抓取网络招聘文本信息。

3.2　岗位抽取

为了从网络招聘文本抽取岗位信息，本文使用 Beautiful Soup 将 HTML 文本转换成树形结构，每个节点对应一个 Python 对象。Beautiful Soup 是一个能从 HTML 或 XML 文件中提取数据的 Python 库。它能通过自定义的解析器来提供导航、搜索，甚至改变解析树。因此，本文使用 Beautiful Soup 获取“岗位”标签内信息。

3.3　知识点抽取

类似于岗位抽取，同样使用 Beautiful Soup 解析网络招聘文本中“任职要求”标签中的文本。然后对文

俞琰，陈磊，赵乃瑄.基于网络招聘文本挖掘的课程知识模型自动构建研究[J].图书情报工作,2019,63(10):134－142.

本进行分词、词性标注、去停用词、英文大小写转换等预处理工作。图5为预处理示例。欲抽取的知识点使用粗体表示。本文使用Python的jieba扩展包进行分词和词性标注，使用哈尔滨工业大学编写的停用词表，过滤除去停用词。

任职要求：1、1年以上 **Java** 开发经验 ；对软件工程和相关标准有良好的认识;具有较强的面向对象思维；精通设计模式；2、熟悉 **Spring**、**MyBatis** 等主流 **J2EE** 技术；熟练使用 **Oracle** 数据库，并有一定的 **SQL** 优化经验；3、熟悉 **Javascript**、**JQuery**、**BootStrap**、**CSS** 等技术；熟悉 **Linux** 操作系统；熟悉 **Tomcat** 应用服务器；4、有 **Spark**、**Hadoop** 开发经验者优先；5、能够承受压力、基础扎实、思路清晰，有独立解决问题的能力、良好的沟通表达能力，有责任心，具有良好的团队合作意识。

预处理

任职 要求： 1、1 年 以上 **java** 开发 经验 ；对 软件工程 和 相关 标准 有 良好 的 认识；具有 较强 的 面向对象 思维；精通 设计模式；2、熟悉 **spring**、**mybatis** 等 主流 **j2ee** 技术；熟练 使用 **oracle** 数据库，并 有 一定 的 **sql** 优化 经验；3、熟悉 **javascript**、**jquery**、**bootstrap**、**css** 等 技术；熟悉 **linux** 操作系统；熟悉 **tomcat** 应用服务器；4、有 **spark**、 **hadoop** 开发 经验 者 优先；5、能够 承受 压力、基础 扎实、思路清晰，有 独立 解决 问题 的 能力、良好 的 沟通 表达能力，有 责任心，具有 良好 的 团队 合作 意识。

图5　预处理示例

为了抽取预处理后文本中的知识点，传统的方法通常使用词频方法，抽取语料集中出现频繁的词作为知识点。然而，基于词频的方法抽取准确率低，常包含"能力""经验"等非知识点词语。知识点具有专业相关性，在某些专业中频繁出现，而在其他专业中很少出现。因此，本文引入包含其他专业集合的辅助集，提出基于辅助集重要性(auxiliary set based importance, ASI)衡量词语在专业的重要性，以抽取知识点。其基本原理是：一个词语在目标集中出现频次越高，在辅助集中出现频次越低，则越可能是目标专业的知识点。

具体地，设待分析的目标集(target set, TS)，包含其他专业招聘信息的辅助集(auxilary set, AS)，衡量一个词语 w_i 在目标集TS中的专业重要性ASI(w_i, TS)定义如下：

$$ASI(w,TS)=\frac{\dfrac{df(w_i,TS)+1}{|TS|}}{\dfrac{df(w_i,AS)+1}{|AS|}} \qquad 公式(1)$$

其中，$df(w_i, TS)$表示在 TS 集合中，包含 w_i 的文本数；$df(w_i, AS)$表示在 AS 集合中，包含 w_i 的文本数；$|TS|$表示 TS 集合中文本数；$|AS|$表示 AS 集合中文本数。由于技能通常为名词，因此本文选出"任职要求"中的名词作为候选词，度量候选词的专业重要性，按专业重要性大小排序，以抽取知识点。

3.4　课程生成

本文使用Latent Dirichlet Allocation (LDA)模型生成隐含的课程。LDA主题模型[23]是自然语言处理中一种常用的三层贝叶斯概率模型。该模型由词、主题和文本三层构成，见图6(a)。模型假设每个文本包含若干隐含主题，每个主题包含特定的词。文本和词间的关系通过隐含主题体现。隐含主题被文本集中所有文本所共享，而每个文本有一个特定的主题分布。一篇文本的构造过程首先是以一定的概率选择某个主题，然后再在这个主题下以一定的概率选出某一个词，这样就生成了这个文本的第一个词。不断重复这个过程，就生成了整个文档。

类似地，满足一个岗位的要求需要学习多门课程，每门课程包含若干个知识点，见图6(b)。因此，本文提出使用LDA主题模型生成隐含的课程信息。

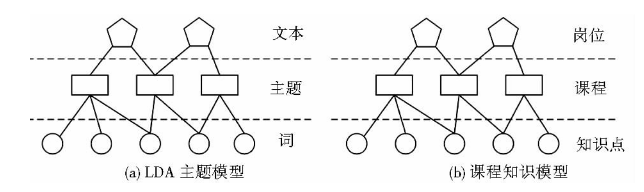


图6　主题模型在课程知识模型的应用

LDA主题模型通常采用Gibbs采样推理方法估计主题的后验分布，计算如公式(2)[24]所示：

$$p(z_{ij}=k \mid z^{-ij}, w, \alpha, \beta) \propto \frac{n_{i(\cdot)k}^{-ij}+\beta}{n_{(\cdot)(\cdot)k}^{-ij}+V\beta} \times \frac{n_{(\cdot)jk}^{-ij}+\alpha}{n_{(\cdot)j(\cdot)}^{-ij}+K\alpha} \qquad 公式(2)$$

其中，z_{ij}表示岗位 d_j 中知识点 w_i 的课程变量；$-ij$ 表示排除岗位 d_j 中的知识点 w_i；n_{ijk}表示岗位 d_j 中的知识点 w_i 分配给课程 z_k 的次数；$(\cdot)$表示对应维度(岗

位、课程、知识点）所有次数之和，β 表示知识点的 Dirichlet 先验分布，α 表示课程的 Dirichlet 先验分布，K 表示课程数，V 表示集合中总的知识点数。一旦获得每个岗位中每个知识点的课程，就可以得到 LDA 模型中 θ 和 φ 的后验估计值，计算如公式（3）[24] 和公式（4）[24] 所示：

$$\theta_{jk} = \frac{n_{(\cdot)jk} + \alpha}{n_{(\cdot)j(\cdot)} + K\alpha} \quad \text{公式（3）}$$

$$\varphi_{ki} = \frac{n_{i(\cdot)k} + \beta}{n_{(\cdot)(\cdot)k} + V\beta} \quad \text{公式（4）}$$

其中，θ_{jk} 表示岗位 d_j 包含课程 z_k 的概率；φ_{ki} 表示课程 z_k 中包含知识点 w_i 的概率。

3.5 岗位－课程关系生成

使用关联性 r 表示岗位 d_i 与课程 z_k 的关系强度，表明岗位 d_i 包含课程 z_k 的平均概率与所有岗位包含课程 z_k 的平均概率比值，其定义如下：

$$r(d_i, z_k) = \frac{\frac{\sum_{d_i \in Di} \theta_{ik}}{|D_i|}}{\frac{\sum_{d_j \in TS} \theta_{jk}}{|TS|}} \quad \text{公式（5）}$$

其中，$D_i = \{d_i | d_i \in TS\}$，表示集合 TS 中所有包含岗位 d_i 的网络招聘文本数量。由公式(5)可知，r 值越大，表明岗位 d_i 和课程 z_k 的关系强度越大；反之，关系强度越小。

3.6 课程－知识点关系生成

课程－知识点关系表明特定课程所包含的主要知识点。由于 LDA 主题模型可以得到 φ_{ki} 表示课程 z_k 中知识点 w_i 的概率，因此为每门课程选择前若干个知识点，生成课程－知识点关系，使用 φ_{ki} 表示课程 z_k 中知识点 w_i 的关系强度。

4 实证

4.1 数据抓取

为了验证本文提出方法的可行性与有效性，实验选择国内主流招聘网站前程无忧（www.51job.com）中的计算机专业本科相关专业进行分析。前程无忧是一家网络招聘服务提供商，是中国最具影响力的人才招聘网站之一。按照职能，在前程无忧网站选取“计算机/互联网/通信/电子”职能抓取数据，数据抓取日期为 2018 年 3 月 19 日至 26 日）。为了得到辅助集，在前程无忧招聘网站依次选取“销售/客服/技术支持”“会计/金融/银行/保险”“生产/营运/采购/物流”“生物/制药/医疗/护理”“广告/市场/媒体/艺术”“建筑/房地产”“人事/行政/高级管理”“服务业”职能抓取数据，数据抓取日期为 2018 年 3 月 19 日至 26 日。抓取后的网页文本去除本科以下学历、内容重复、全英文、没有写明任职要求的招聘文本，最后得到的数据基本信息如表 1 所示：

表 1 数据集基本信息

数据集类型	专业	网络招聘文本数(篇)
目标集	计算机/互联网/通信/电子	14 678
辅助集	销售/客服/技术支持	2 361
	会计/金融/银行/保险	2 417
	生产/营运/采购/物流	2 303
	生物/制药/医疗/护理	2 257
	广告/市场/媒体/艺术	2 578
	建筑/房地产	2 343
	人事/行政/高级管理	2 269
	服务业	2 373
	总计	18 901

4.2 岗位抽取

通过招聘网页岗位名称中词语频次统计，去除“开发”“研发”“工程师”等不能表示明确岗位的词后，排在前 10 的高词频词形成的岗位，以及本文给出的标准化岗位名称，结果如表 2 所示。由表 2 可见，计算机学科技术更替非常快。虽然存在一些持续热门的岗位，如 Java 工程师、C++开发工程师、.net 工程师等，但是也有一些新的岗位需求量快速增长，如前端开发工程师、大数据工程师、算法工程师等。

表 2 计算机相关专业前 10 岗位

序号	高频词	包含关键词的岗位	标准化岗位名称
1	Java	Java 软件工程师、Java 工程师、Java 开发工程师	Java 工程师
2	数据	大数据工程师、大数据研发工程师、大数据研发人员	大数据工程师
3	C++	C++开发工程师、C++软件工程师、C++软件开发工程师	C++开发工程师
4	.net	.net 开发工程师、.net 工程师、.net 软件开发工程师	.net 工程师
4	前端开发	Web 前端开发工程师、前端开发工程师	前端开发工程师
5	测试	测试工程师、软件测试工程师	测试工程师
6	运维	运维工程师、系统运维工程师	运维工程师
8	嵌入式软件	嵌入式软件工程师、嵌入式软件开发工程师	嵌入式软件工程师
9	IOS	IOS 开发工程师、IOS 移动研发工程师	IOS 开发工程师
10	算法	算法工程师、人工智能算法工程师、AI 算法工程师、图像处理算法工程师、自然语言处理算法工程师	算法工程师

4.3 知识点抽取

知识点抽取选取 ASI 值排序生成,采用人工方式进行判断。为了避免主观性和专业知识的局限性,利用百度百科、维基、互动百科等知识网站查找是否存在对应的知识点词条,以判别被抽取知识点的正确性。表 3 给出使用词频 TF 方法与本文提出方法抽取的前 10 个词及对应值,其中粗体词语表示非知识点词。由表 3 可见,TF 前 10 个词语中,非知识点词因在目标集中出现频次高,不能很好地抽取目标集中的知识点。而使用 ASI 方法识别的前 10 个词语均为技能,明显优于 TF 方法,因为"经验""能力"等目标集中的高频词在辅助集中也高频出现,使得这些词的 ASI 值变小。

表 3 不同方法识别的前 10 个词比较

序号	TF	TF 值	ASI	ASI 值
1	经验	32 622	Java	3 950.752
2	能力	27 070	C ++	2 442.615
3	技术	23 681	MySQL	2 194.096
4	职责	22 823	SQL	2 054.952
5	学历	22 742	JavaScript	1 502.328
6	岗位	22 738	Python	1 178.912
7	专业	22 497	JQuery	1 063.164
8	数据	22 007	C#	951.500
9	软件	21 231	面向对象	910.648
10	计算机	21 068	Ajax	781.282

4.4 课程生成

为了生成课程,使用 LDA 模型,依据常见参数设置方法[23－24],主题模型设置 $\alpha=50/K$、$\beta=0.01$,Gibbs 采样迭代次数参数为 2 000,保存迭代参数为 1 000。课程数 K 的选取通过计算困惑度与专家评估选取最优值,采用五折交叉验证。根据计算,实验设定课程数 K =11。表 4 列出各个课程前 5 个知识点以及归纳的对应课程名。

表 4 课程名生成

序号	知识点	课程名
1	Web HTML JavaScript CSS HTML5	Web 开发
2	Java J2EE Spring 框架 Hibernate	Java
3	C#. net Winform 面向对象 软件架构	C#
4	编程 C ++ C Linux Unix	C ++
5	数据库 Oracle SQL MySQL 存储过程	数据库
6	数据分析 MapReduce 存储 建模 数据挖掘	数据分析
7	Linux 底层 进程 Shell 通信	Linux
8	协议 TCP IP HTTP 通信	网络通信
9	软件测试 Bug 测试用例 单元测试 白盒	软件测试
10	软件工程 设计模式 架构设计 重构 敏捷	软件工程
11	算法 Python C ++ 视觉 AI	人工智能

在这 11 门课程中,一些课程是许多高校开设多年的计算机专业课程。如,"Java""C ++""C#""数据库""Linux""网络通信""软件测试"与"软件工程"等。也有一些课程是随着大数据而新出现的课程,如"Web 开发""数据分析"等。

4.5 岗位－课程关系生成

根据岗位和课程关联度计算,得到岗位－课程之间的关系强度,表 5 列出岗位与课程的关系。

表 5 岗位－课程关系

课程	岗位									
	Java 工程师	大数据工程师	C ++ 开发工程师	. net 工程师	前端开发工程师	测试工程师	运维工程师	嵌入式软件工程师	IOS 开发工程师	算法工程师
WEB 开发	★★			★★	★★★				★	
Java	★★★									
C/C ++			★★★			★		★★★	★★★	★★
C#				★★★						
数据库	★★	★★	★★	★★	★★	★★	★★★		★	
数据分析		★★★								★★
Linux			★★★			★	★★★	★★★	★★	★
网络通信			★★				★	★★	★★	
软件工程	★★			★★						
软件测试						★★★				
人工智能		★								★★★

注:★表示岗位－课程关联度 r [0.9,1),★★表示岗位－课程关联度 r [1, 1.5),★★★表示岗位－课程关联度 r [1.5,3)

通过表 5 的结果,可以看出各岗位所需学习的主要课程:"Java 工程师"运用 Java 开发语言去完成软件产品的程序设计、开发等工作,主要从事负责运营平台核心后台业务及时对外服务接口的设计与开发。通常

需要学习 Java、J2EE 框架、数据库、前端开发、软件工程等课程。

“大数据工程师”使用现代数据仓库技术、线上分析处理技术、数据挖掘和数据展现技术进行数据分析以实现商业价值。因此，除了需要掌握传统的数据库技术之外，大数据工程师需要熟悉分布式数据存储、分布式计算和数据挖掘的原理。

“C ++ 开发工程师”主要从事 Windows 或 Linux 平台下 C ++ 软件编程，主要需要掌握 C ++ 、Linux 操作系统、网络通信和数据库等课程。

“. net 开发工程师”利用微软的. net 开发 Web 程序、Windows 应用程序和 Wap 无线网络应用程序等。. net 开发工程师主要需要学习 C#、数据库、前端开发、软件工程等课程。

“Web 前端开发工程师”是一个很新的职业，主要进行网站开发、优化、完善的工作。一位合格的 Web 前端开发工程师首先需要掌握前端开发的各门课程，此外，还需要熟悉传统的数据库知识、面向对象等软件工程知识。

“测试工程师”：我国的软件测试职业还处于一个发展的阶段，很多中大型软件企业设立了单独的测试部，与开发部并行运作。作为一名测试工程师，需要掌握主要的测试原理和工具，还需要熟悉主流的操作系统和数据库。

“运维工程师”主要负责维护并确保整个服务的高可用性，同时不断优化系统架构提升部署效率、优化资源利用率。运维工程师面对的最大挑战是大规模集群管理问题，因此运维工程师主要需要掌握操作系统、网络通信以及数据库。

“嵌入式软件工程师”是编写嵌入式系统的工程师。嵌入式系统是以应用为中心，以计算机技术为基础，并且软硬件可裁剪，适用于应用系统对功能、可靠性、成本、体积、功耗有严格要求的专用计算机系统。嵌入式软件工程师主要需要掌握 C ++ 、Linux、网络通讯等技能。

“IOS 开发工程师”主要以 IOS 系统为基础的手机等便携终端为基础，进行相应的开发工作。该岗位主要需要掌握 C ++ 、Linux、网络通信、前端开发等课程。“算法工程师”主要研究包括机器人、语言识别、图像识别、自然语言处理和专家系统等，从大量的数据中通过算法搜索隐藏于其中的知识。该岗位主要需要掌握人工智能原理与算法、数据分析、操作系统以及 C ++ 等课程。

4.6 课程 – 知识点关系生成

使用 LDA 主题模型，得到 11 门课程与知识点之间的概率关系，选取每门课程的前 15 个知识点词语，构成课程 – 知识点关系，形成词云，如图 7 所示：



图 7 课程知识点词云

由图 7 可见，目前高校普遍开设的计算机专业课程需要关注市场的新需求，添加新的知识点。如：课程“Web 开发”早期网站内容开发主要是静态的、以图片和文字为主。随着互联网技术的发展和 HTML5、CSS3 等技术和框架的引入，现代网页更加美观，功能更加强大，所以课程需要强化这些新技术的学习。课程“Java”是一门面向对象编程语言，J2EE 是一个为大企业主机级的计算类型而设计的 Java 平台，简化了应用程序的开发，也降低了对编程的要求，因此课程需要加强 J2EE 以及相关框架的学习，以满足企业的需要。由于企业进行系统开发的敏捷性与代码的可维护性，需要涉及一些架构，所以课程“C#”需要加强软件架构和设计模式的学习。课程“C ++ ”除了学习语言本身的语法知识之外，也需要注重其在 Linux、Unix 系统上的应用与开发。课程“数据库”是管理信息系统、办公自动化系统、决策支持系统等各类信息系统的核心部分，是进行科学研究和决策管理的重要技术手段。近年来，随着数据量的高速增长，分布式数据库技术快速发展。传统的关系型数据库开始从集中式模型向分布式架构发展，以 NoSQL、MongDB 为代表的非关系型数据库，因其高可扩展性、高并发性等优势而快速发展。在教师授课过程中，需要密切关注这些非关系型数据库的发展趋势与介绍。课程“人工智能”是计算机科学的一个分支，它企图了解智能的实质，并生产出一种新的、

能以人类智能相似的方式做出反应的智能机器,目前的教学与学习需要关注最近企业的热点应用,如:机器人、语言识别、图像识别、自然语言处理和专家系统等知识点。

随着大数据互联网的飞速发展,也出现了一些新兴课程,课程知识模型也为这些新出现的课程教学大纲和知识点的设置提供了情报决策的依据。如:课程“数据分析”将组织透过咨询系统之联机事务处理经年累月所积累的大量资料,透过数据仓库理论所特有的资料存储架构,通过 Spark、Hadoop 大数据集群计算环境,作系统地分析整理。利用各种分析方法,如数据挖掘,进而支持决策支持系统的创建,帮助决策者快速有效地从大量资料中分析出有价值的咨询。以利决策拟定及快速回应外在环境变动,帮助构建商业智能。数据分析、Spark、Hadoop、MapReduce 等均是本课程设置需要考虑的知识点。

5 结语

目前研究主要对网络招聘文本中的岗位所需技能知识点进行手工分析,没有进一步利用网络招聘文本信息。针对目前网络招聘信息分析存在的问题,本文提出包含“岗位－课程－知识点”的三级课程知识模型,并通过自然语言处理、文本挖掘技术实现了课程知识模型的自动构建,最后对计算机相关专业的网络文本进行实证分析。实证结果表明了本模型以及构建过程的可行性与有效性。通过分析,可以发现企业这些岗位对人才专业技能的主要需求,为高等院校专业设置、教师教学大纲知识点设置、学生职业规划和知识点补充起到指导性作用,从而缓解找工作难、招聘难的双重矛盾。

由于岗位名称的多样性,如:“Java 开发工程师”“Java 软件工程师”均表示相同的含义,目前的研究方法主要采用主要关键词“Java”标准化为相同岗位名称,后续的研究中将进一步优化岗位名称标准化的方法,以自动、准确地表示岗位信息。

参考文献:

[1] CHAO C, SHIH S. Organizational and end-user information systems job market: an analysis of job types and skill requirements[J]. Inform techno learn perform, 2005, 23(1): 1－15.

[2] WOWCZKO I. Skills and vacancy analysis with data mining techniques[J]. Informatics, 2015, 2(4):31－49.

[3] KIM J Y, LEE C K. An empirical analysis of requirements for data scientists using online job postings [J]. International journal of software engineering and its application, 2016, 10(4): 161－172.

[4] MAURO D A, GRECO M, GRIMALDI M, et al. Beyond data scientists: a review of big data skills and job families[C]//Proceedings of the 2016 international forum on knowledge asset dynamics. Berlin: Springer International Publishing, 2016: 1844－1857.

[5] 吕斌,张通,周珏.面向组织的具有通用性的情报职业及情报从业人员——基于组织招聘网页信息挖掘的分析之一[J].图书情报工作, 2009, 53(4): 19－23.

[6] 李国秋,桑培铭.情报过程——情报职业的核心:问题域及方法论——基于组织招聘网页信息挖掘的分析之二[J].图书情报工作, 2009, 53(4): 24－27.

[7] 夏火松, 潘筱听. 基于 Python 挖掘的大数据学术研究与人才需求的关系研究[J]. 信息资源管理学报, 2017, 7(1): 4－12.

[8] 黄崑, 王凯飞, 王珊珊,等. 数据类岗位招聘需求调查及对图情学科人才培养的启示[J]. 图书情报知识, 2016 (6):42－53.

[9] 贾东琴, 檀博. 国外高校图书馆岗位需求调研分析——基于招聘广告的内容分析[J]. 图书馆建设, 2018 (2): 84－89.

[10] SODHI M, SON B. Content analysis of OR job advertisements to infer required skills[J]. The journal of the Operational Research Society, 2010, 9(1): 1315－1327.

[11] ZHAO M, JAVED F, JACOB F, et al. SKILL: a system for skill identification and normalization[C]// Proceedings of the twenty-seventh conference on innovative applications of artificial intelligence. Palo Alto: AAAI, 2015: 4012－4017.

[12] XU T, ZHU H, ZHU C, et al. Measuring the popularity of job skills in recruitment market: a multi-criteria approach[C]//Proceedings of the 32nd AAAI conference on artificial intelligence, Menlo Park: AAAI, 2018: 3013－3028.

[13] 詹川. 基于文本挖掘的专业人才技能需求分析——以电子商务专业为例[J]. 图书馆论坛, 2017, 5(1): 116－123.

[14] 夏立新, 楚林, 王忠义,等. 基于网络文本挖掘的就业知识需求关系构建[J]. 图书情报知识, 2016, 169(1):94－100.

[15] BASTIAN M, HAYES M, VAUGHAN W, et al. LinkedIn skills: large-scale topic extraction and inference[C]// ACM conference on recommender systems. New York: ACM, 2014:1－8.

[16] 王召义, 薛晨杰, 刘玉林. 基于邻近词分析的电子商务技能需求分析[J]. 信息资源管理学报, 2018, 11(2): 113－121.

[17] 刘睿伦, 叶文豪, 高瑞卿,等. 基于大数据岗位需求的文本聚类研究[J]. 数据分析与知识发现, 2017, 12(12): 32－40.

[18] 张俊峰, 魏瑞斌. 国内招聘类网站的数据类岗位人才需求特征挖掘[J]. 情报杂志, 2018, 37(6): 176－182.

[19] 王东波, 胡昊天, 周鑫,等. 基于深度学习的数据科学招聘实体自动抽取及分析研究[J]. 图书情报工作, 2018, 62(13): 64－73.

[20] 魏来, 郑华敏. 国内外数据馆员能力要求比较研究[J]. 图书情报工作, 2018,62(10): 18－24.

[21] 田野. 国内图情档专业需求现状调查与分析[J]. 图书情报工

作, 2018, 62(9): 62 - 72.
[22] 陈媛媛, 董伟. 社会需求导向下图书情报专业毕业生就业技能分析[J]. 图书情报工作, 2017, 61(19):66 - 73.
[23] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet Allocation [J]. Journal of machine learning research, 2003, 3(1):993 - 1022.
[24] GRIFFITHS T L, STEYVERS M. Finding scientific topics [J]. PNAS, 2004, 101(1): 5228 - 5235.

作者贡献说明:

俞琰:提出研究思路,设计研究方案,进行数据建模,撰写论文;

陈磊:进行数据收集和数据清洗;

赵乃瑄:修改论文。

Research on Automatic Construction of Curriculum Knowledge Model Based on Web Recruitment Text Mining

Yu Yan[1,2] Chen Lei[1] Zhao Naixuan[1]

[1] Information Service Department, Nanjing Tech University, Nanjing 210009

[2] Computer Science Department, Southeast University Chengxian College, Nanjing 211816

Abstract: [**Purpose/significance**] In order to help college teachers and students make full use of web recruitment information, this paper proposes a curriculum knowledge model and its automatic construction method based on large data web recruitment text mining. [**Method/process**] This paper proposes a three-level curriculum knowledge model including "post-curriculum-knowledge point", which uses natural language text mining technology to realize the automatic construction, and verifies the construction process through experiments. [**Result/conclusion**] The experimental results show that the proposed model and method are highly feasible and effective, and provide teaching and learning reference for colleges and students.

Keywords: Web recruitment text curriculum knowledge model text mining